Hybrid SORS: Breaking Through the ML/AI Computational Barrier

Objectives

Abstract: It is a surprise to many, but the frenetic activity spawned in ML and AI over the past decade has not been driven by theoretical advances. The fundamental underpinnings of current ML/AI approaches were published by Arthur Samuelson in 1952, and while there have certainly been improvements, the approach is basically the same. What has instead driven the activity has been available computational power. In 1952, computers would require months to evaluate a simple network. Over the last couple of decades, computational power has reached the point where interesting networks can be evaluated in useful time. "Useful" time is an interesting concept, but it is clear from the number of startups (and failed startups) striving to develop higher-performance lower-power AI accelerators that VCs and entrepreneurs believe that real future for AI involves achieving another 10-100x acceleration. Whether that speedup is possible is an open question, but there are some definitive statements that that can be made on the question:
- The problems that must be solved to achieve this speed up are not new problems. "Many have tried; none have succeeded."
- The solution, if there is one, will involve parallelism.
- The solution, if there is one, will not be the development of a new, "magical" parallel hardware architecture.
- Following the lessons of Linpack versus LAPACK, data reuse will be a key part of the solution.
- Following lessons garnered from the vector world, compiler-managed memory (i.e. registers) will almost certainly be a key part of the solution.
This talk will discuss these points and what they suggest about a possible solution to this problem. Not surprisingly, compilers will be a critical tenet of the solution.

Short bio: Being a firm believer

in the statement “Implementations are oft-needed checks on the egos of theoreticians”, Randy Allen’s has likewise centered his career on developing both theory and implementation of innovative, highly optimizing tools and software. Dr. Allen’s PhD dissertation focused on optimizing compilers for parallel and vector machines, culminating in the graduate-level textbook (coauthored with Ken Kennedy) “Optimizing Compilers for Modern Architectures”. As VP of Performance Engineering at Chronologic Simulation, Randy was an early developer of VCS, the world’s first compiled Verilog simulator. Catalytic, Inc (founded by Randy), pioneered fixed-point MATLAB(r) compilers for DSP’s. Dr. Allen’s current interests are developing high-performance, low-power architectures and compilers that can meet the needs of ML/AI networks. Dr. Allen earned an A.B. summa cum laude in Chemistry from Harvard University and a PhD in Mathematical Sciences from Rice University.

Speakers

Speaker: Randy Allen, CEO Resonant Solutions
Host: John Davis, LOCA director and EEA Group Leader, Computer Sciences

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---